Improved Multitaper PNCC Feature for Robust Speaker Verification

Liu Yi, He Liang, Liu Jia
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
Email: liuyi12@mails.tsinghua.edu.cn, {heliang, liuj}@tsinghua.edu.cn

***Abstract*—A major challenge for practical speaker verification is the significant performance degradation in noisy circumstances. Recent works have shown that some modern features are promising to improve the system robustness. This paper introduces a novel feature based on power normalized cepstral coefficients (PNCC), improved by multitaper and compressive gammachirp filter-bank (cGCFB). Our analysis indicates that the improved multitaper PNCC (imPNCC) benefits from better spectrum estimation, and thus outperforms competitors with a proper configuration. The proposed method shows 4.28% and 3.25% relative improvement in EER over the baseline PNCC feature on NIST 2008 SRE telephone-telephone female condition using state-of-the-art i-vector speaker verification system.**

***Keywords*—Robust feature; speaker verification; imPNCC; multitaper method; cGCFB.**

I. Introduction

Speaker verification, which is intended to verify whether a given utterance is spoken by a specific speaker, has been developed by leaps and bounds during last decades. Gaussian mixture model and support vector machines (GMM-SVM), joint factor analysis (JFA), i-vector followed with probabilistic linear discriminant analysis (PLDA) are representatives of these technology [1] [2].

Even though the performance in the laboratory condition using i-vector system is satisfying, it exhibits a sharp deterioration in the presence of various noise and interference. The decline in performance is due to the mismatch between training and test condition, the disturbance including noise, reverberation, etc. As a major contributor, background noise generally exists in practical applications, which must be considered to design robust speaker verification. Plenty of algorithms have been proposed trying to address this problem. One category is to suppress noise from the perspective of signal processing. Vector Taylor series (VTS) [3] and ETSI advanced front end (AFE) [4] belong to this category. The disadvantage of these methods is that they require much more computation and often fail to work in front of complex environments. Meanwhile, model level compensation is also presented, but suffers the same problem(fails for the same reason).

Recently, results reveal that modern robust features are promising under non-stationary noise environment [5]. Normalized modulation cepstral coefficients (NMCC) is proposed in [6]. NMCC treats speech as a combination of amplitude modulation (AM) signals and extracts cepstral coefficients from sub-band AM power spectrum. Another robust feature, presented by Sadjadi and Hansen, is mean Hilbert envelope coefficients (MHEC) [7]. The feature is based on the Hilbert envelope of gammatone filter-bank outputs. A new asymmetric noise suppression filter is designed in power normalized cepstral coefficients (PNCC) feature which is invented by Kim et al. [8], to directly estimate the level of background noise and remove it from spectrum.

It has been demonstrated that above three features improve acoustical system robustness. Furthermore, some experiments indicate that PNCC achieves an overall superiority in performance [9]. Nevertheless, as no noise is involved, PNCC becomes less powerful than traditional features like MFCC or PLP [9] [10]. To solve the problem, Two methods are proposed in this paper. First, inspire by [11], multitaper method is embedded with PNCC for the first time. Multitaper method makes a tradeoff between bias and variance of spectrum estimation, and a configuration that is more suitable for speaker verification could be selected. Second, compressive gammachirp filter-bank (cGCFB), which is an alternative to the widely used gammatone filter-bank, is involved. The cGCFB is expected to enhance the robustness of sub-band power integration. The proposed feature, named improved multitaper PNCC (imPNCC), aims to increase the performance of PNCC on both clean and noisy condition, without much more additional computational overhead.

The outline of the paper is as follows. We describe the standard procedure of PNCC feature extraction in Section 2. Section 3 presents the imPNCC. The experiment setup(?) and results are given in Section 4. Section 5 concludes the paper.

II. Power Normalized Cepstral Coefficients

The structure of PNCC feature extractor is similar with conventional MFCC and PLP. The procedure is briefly reviewed below. For readers who want more details, please refer to [12]. Note that we assume the sampling rate to be 8KHz.

As other acoustic features, a pre-emphasis filter of the form $H(z) = 1 - 0.97z^{-1}$ is applied first. Then a short-time Fourier transform (STFT) using Hamming windows with duration 25 ms and shift length 10 ms, is performed. The squared magnitudes of STFT output are summed up to n sub-band power $P[m, l]$ specified by a gammatone filter-bank. Here, m is the frame index and $1 \leq l \leq n$ is the channel index. The center frequencies of the filter-bank are linearly spaced in equivalent rectangular bandwidth domain between f_{low} and f_{high}.

The integrated outputs of a gammatone filter-bank is averaged by medium-duration windows (typically 65 ms), and

treated as inputs of asymmetric noise suppression (ANS) filter S_i. The ANS filter explicitly estimates noise level based on the fact that the spectrum associated with most noise changes more slowly than the instantaneous power of human speech. The envelope of the smoothed power is evaluated by a rather simple low-pass filter expressed by

$$S_o[m,l] = \begin{cases} \lambda_a S_o[m-1,l] + (1-\lambda_a) S_i[m,l], \\ \quad \text{if } S_i[m,l] \geq S_o[m-1,l] \\ \lambda_b S_o[m-1,l] + (1-\lambda_b) S_i[m,l], \\ \quad \text{if } S_i[m,l] \leq S_o[m-1,1] \end{cases} \tag{1}$$

where S_o denotes corresponding output of the filter and if $0 < \lambda_b < \lambda_a < 1$, S_o tends to follow the lower envelop of the input S_i. The output is served as a model of averaged noise power N and the power above this envelop is considered to represent speech activity [12]. Temporal masking and spectral weight smoothing are followed to derive a weighting function from S_o to modulate the original filter-bank power.

The results are once again normalized by a running average of the overall power to minimize the impact of amplitude scaling. Instead of intrinsic logarithmic or cube-root , a power-law function with an exponent of 1/15 is chosen. The function is carefully designed for fitting the curve relating sound pressure level to the human auditory-nerve firing rate. Discrete cosine transform (DCT) and cepstral mean normalization are final stages of the extraction so that 13-dimension feature is obtained.

III. Methodology to Improve Original PNCC

Although PNCC feature shows an advantage under noise environment, it shows no advantages over MFCC or PLP when the background is clean. As previous studies claimed, a gammatone filter-bank provides robustness for acoustic system [8] [13]. Empirically, a better sub-band power estimation is generally promising to increase the recognition accuracy. An improved multitaper PNCC (imPNCC) feature proposed in this paper originated from this idea and explores two front-end methods to improve the performance of PNCC. The flowchart of imPNCC feature extractor is shown in Fig. 1. The shaded blocks emphasize the main differences between imPNCC and PNCC.

A. Low-variance multitaper method

Given a segment of a utterance, $\mathbf{x} = [x(0), x(1), \ldots, x(L-1)]$ is one frame in time domain with length L. The typical short-time spectrum estimate $\hat{P}(f)$ utilizes windowed short-time fourier transform, which is expressed as

$$\hat{P}(f) = \left| \sum_{t=0}^{L-1} w(t)x(t) \exp(\frac{-j2\pi tf}{L}) \right|^2 \tag{2}$$

where j is the imaginary unit, $f = 0, 1, \ldots, N-1$ denotes the discrete frequency, and $w(t)$ is a window function and the most popular one is Hamming window.

Multitaper method is distinguished by the usage of multiple orthogonal windows, called tapers. Weighted averaging in frequency-domain is followed to obtain the final spectrum estimation. The process is given by

$$\hat{P}(f) = \sum_{n=1}^{K} \lambda(n) \left| \sum_{t=0}^{L-1} w_n(t)x(t) \exp(\frac{-j2\pi tf}{L}) \right|^2 \tag{3}$$

Here $w(t)$ is replaced by K tapers $w_n(t)$ with weight $\lambda(n)$. Notice that, if $K = 1$, $\lambda = 1$, (3) simply degrades to (2).

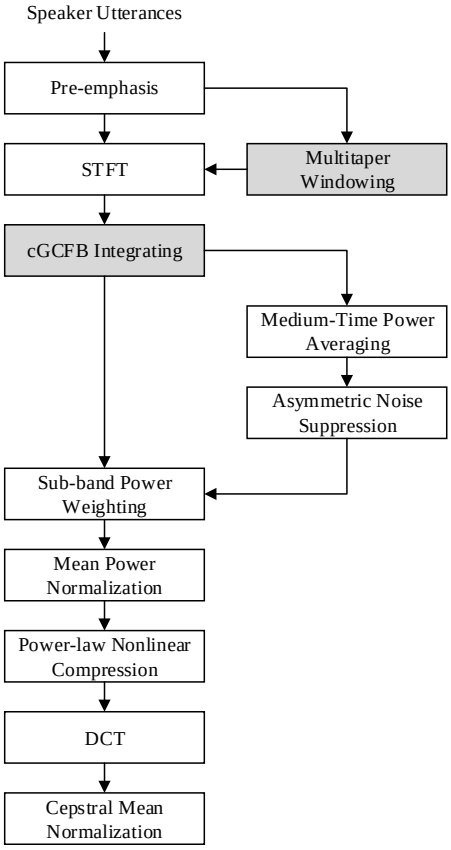


Fig. 1. Flowchart of imPNCC feature extractor. The difference from PNCC is emphasized by as shaded blocks.

Although multitaper technology has been proposed for a long time and some of its applications in speech signal processing have been reported [14], it does not receive much attention in feature extraction until Kinnunen *et al.* first investigated the use of multitaper in robust speaker verification [11]. It is claimed that a multitaper spectrum possesses a smaller variance than the single-windowed estimation and is more powerful especially when the spectrum of interest varies rapidly. The method significantly improves the performance of state-of-the-art i-vector speaker verification system using MFCC and PLP as features [15].

Multitaper is introduced to imPNCC and three types of tapers, i.e. Thomson [16], sine [17], multipeak [18], are studied in this paper. We denote imPNCC using the three tapers as imPNCC-t, imPNCC-s and imPNCC-m respectively. As an important parameter, the number of tapers K affects the performance, and should be determined by further experiments. Since a multitaper estimator requires K STFT, the computational complexity is higher than single-windowed method.

However, if the taper coefficients are calculated offline, additional computation is quite limited, compared to the entire imPNCC extraction process.

B. Compressive gammachirp auditory filter-bank

Suitable perceptual filter-banks such as gammatone provide robustness to acoustic spectral features. The gammatone filter-bank, which is expected to simulate human cochlear filtering, has been used in features for a long time. However, the gammatone filter is symmetric in frequency and loses ability to capture human auditory perception response as stimulus level increases. The compressive gammachirp filter-bank (cGCFB) was developed to extend the domain of conventional gammatone [19]. cGCFB provides controllable asymmetry to simulate the auditory filter in a more precise way, which make it an ideal alternative to gammatone.

The frequency response of cGCFB is defined as

$$|G_{cGC}| = |G_{GT}| \cdot \exp(c_1\theta_1) \cdot \exp(c_2\theta_2) \tag{4}$$

where $|G_{GT}|$ is the spectral magnitude of gammatone filter, $c_1 = -2.96$ and $c_2 = 2.20$ are fixed parameters given in [19], $\exp(c_1\theta_1)$ and $\exp(c_2\theta_2)$ denote separate low-pass and high-pass asymmetric functions, and θ_i is given by

$$\theta_i = \arctan\left(\frac{f - f_{ri}}{b_i \,\mathrm{ERB}(f_{ri})}\right), \quad i = 1, 2 \tag{5}$$

Again, $b_1 = 1.81$ and $b_2 = 2.17$ are constants. Here, f_{r1} and f_{r2} are center frequencies of corresponding low- and high-pass functions. And f_{r1} shares the same values with the center frequencies of gammatone filter-bank. Furthermore, the relationship between f_{r1} and f_{r2} is given by [19]

$$f_{p1} = f_{r1} + c_1 b_1 \mathrm{ERB}(f_{r1})/n_1 \tag{6}$$

$$f_{r2} = f_{rat} \cdot f_{p1} \tag{7}$$

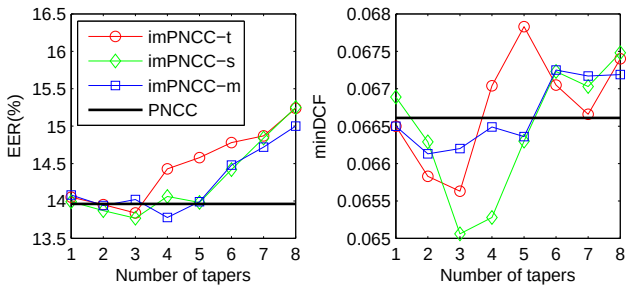
where f_{rat} is the frequency ratio representing the level dependency.

From (4), we see that cGCFB can be conveniently obtained from original gammatone filter-bank by multiplying two specified passband filters. The cGCFB is used in our method to replace gammatone in PNCC.

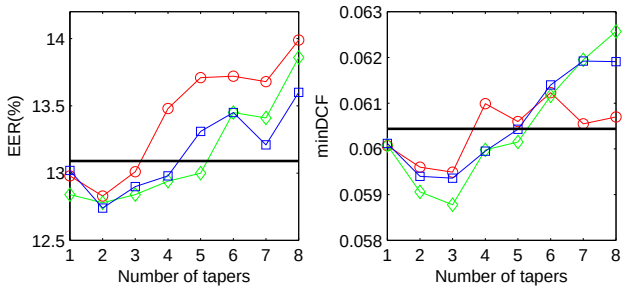
IV. Experimental Setup and Results

In this section, we first figure out the proper number of tapers for imPNCC, and then compare its performance with other features under i-vector based speaker verification. We mainly focus on three features: PLP, PNCC, and imPNCC. MFCC is neglected for the similar performance to PLP. The parameters of PNCC is set followed the instruction in [12] and imPNCC shares totally the same values expect for the different front-end. 13-dimension static features are selected and warped using a 3-seconds sliding window. Delta and delta-delta coefficients are calculated within 5 frames to produce 39-dimension features. imPNCC-t, imPNCC-s, imPNCC-m are evaluated individually. When tuning the number of tapers, cGCFB is excluded from imPNCC.

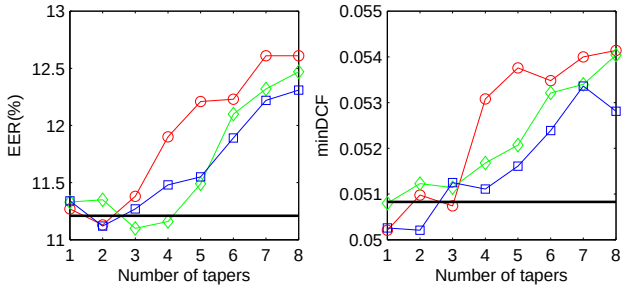
In order to investigate the robustness of these features, we manually mix white noise sampled from NOISEX-92 database [20] with test utterances at different signal-to-noise ratio (SNR). The performance is evaluated by equal error rate (EER) and minimum detection cost function defined by the NIST 2008 SRE evaluation plan (DCF08).



(a) SNR=15dB



(b) SNR=20dB



(c) clean speech

Fig. 2. Effects of different numbers of tapers under 15dB, 20dB and clean speech

A. Tuning the number of tapers

The process of PNCC feature extraction is different from PLP. Therefore, the previous conclusion that the optimal number of tapers K ranges between $3 \leq K \leq 8$ is no longer suitable for imPNCC. The best value of K in our case should be rediscovered. For simplicity and speed, we choose the GMM-UBM system and evaluate the performance on NIST 2008 SRE short2-short3 tel-tel female condition. The database used to train a 256-mixture UBM is NIST 2004 SRE 1-side female corpora. The SNR is sampled at 15dB, 20dB and infinite (no noise involved) to compare the performance under different tapers.

From Fig. 2, it shows that with a proper number of tapers, imPNCC leads to better results than PNCC. Besides, in our case, the optimal K for imPNCC differs from [11] and lies within $2 \leq K \leq 4$, significantly fewer than PLP.

TABLE I. COMPARISON OF RESULTS BETWEEN PLP, PNCC AND IMPNCC ON DIFFERENT TRIAL CONDITION

	5dB		10dB		clean	
NIST 2008 SRE	EER	DCF08	EER	DCF08	EER	DCF08
PLP	12.70	0.0570	9.66	0.0465	5.92	**0.0303**
PNCC	9.68	0.0457	8.17	0.0415	6.02	0.0313
imPNCC-t	9.40	0.0470	8.08	0.0408	**5.82**	0.0313
imPNCC-s	9.36	**0.0453**	8.16	0.0410	**5.82**	0.0307
imPNCC-m	**9.24**	0.0458	**7.82**	**0.0406**	5.87	0.0313
NIST 2010 SRE	EER	DCF08	EER	DCF08	EER	DCF08
PLP	19.64	0.0903	13.31	0.0688	**4.22**	**0.0189**
PNCC	12.19	**0.0633**	7.71	0.0404	4.36	0.0203
imPNCC-t	13.00	0.0667	8.67	0.0426	4.38	0.0198
imPNCC-s	12.93	0.6747	8.56	0.0425	4.57	0.0204
imPNCC-m	**11.92**	0.0635	**7.47**	**0.0396**	4.28	0.0198

Unlike PLP, PNCC is more dependent on the explicit envelop estimation of input speech. Too large value of K would make the spectrum estimation too much smoother [15]. The smoothness blurs the change of spectrum in adjacent channels and frames, and thus impacts the envelop estimation. The blurring effects become more severer when the speech is relatively clean. This possible reason causes the degradation of the performance with a large K. Therefore, the choice of K in imPNCC is critical.

The best results for imPNCC-t, imPNCC-s and imPNCC-m are obtained when K is $2, 3, 2$, respectively. In the next experiment, we fix K and focus on the i-vector system using imPNCC feature.

B. Comparison with other features

To fully recognize the performance of imPNCC feature, i-vector based speaker verification system was used. The experiment were carried out on the female part of the NIST 2008 core condition-6 (tel-tel) and NIST 2010 core-extended condition-1 (int-int) SRE data. The gender-dependent 1024-mixture UBM is trained by randomly selecting 4000 utterances from NIST 2004, 2005, 2006 SRE and Switchboard corpora. A 400-dimension gender-dependent total-variability space and G-PLDA parameters are trained on the same database, but with full collection. The result is shown in Table I.

Table I compares the the results of i-vector classifier using different features. When SNR is low, PNCC gives obviously better results than PLP, as many researches claimed. However, it should be noticed that while no noise is involved, there exists a performance gap between PNCC and PLP. On the other hand, among the three imPNCC features, multipeak shows overall advantages over Thomson and sine tapers. The imPNCC-m feature achieves similar performance to PLP under clean condition. Meanwhile, it outperforms PNCC as the SNR reduces. The results confirm the effectiveness and robustness of imPNCC.

V. CONCLUSION

In this paper, a robust feature named imPNCC is proposed based on PNCC and several experiments have been conducted. We find that when the number of tapers is carefully chosen, the new feature improves the performance while maintains the advantage of PNCC. By virtue of multitaper method and cGCFB, imPNCC achieves similar performance to PLP in clear speech, and improves robustness of PNCC simultaneously. Future work includes fully exploration of the validation of imPNCC on different trial conditions, associated with other types of noise, for example, babble and street noise. It is also interesting to investigate the effectiveness of these two methods in other robust features.

ACKNOWLEDGMENT

The work was supported by National Natural Science Foundation of China (Project No. 61370034 and No. 61273268).

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[3] P. Moreno, B. Raj, and R. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, May 1996, pp. 733–736 vol. 2.

[4] A. Agarwal, , A. Agarwal, and Y. M. Cheng, "Two-stage mel-warped wiener filter for robust speech recognition," in *Proc. ASRU*, 1999, pp. 12–15.

[5] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, Sept 2005.

[6] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, March 2012, pp. 4117–4120.

[7] S. Sadjadi and J. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, May 2011, pp. 5448–5451.

[8] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. INTERSPEECH*, 2009.

[9] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for nist 2012 speaker recognition evaluation," Aug 2013. [Online]. Available: http://www.sri.com/sites/default/files/publications/intspch_noise-robust_l.ferrer_final.pdf

[10] L. Fan, D. Ke, X. Fu, S. Lu, and B. Xu, "Power-normalized plp (pnplp) feature for robust speech recognition," in *Proc. Chinese Spoken Language Processing*, Dec 2012, pp. 224–228.

[11] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-variance multitaper MFCC features: A case study in robust speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 1990–2001, Sept 2012.

[12] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process*, (accepted). [Online]. Available: http://www.cs.cmu.edu/~chanwook/MyPapers/OnlinePNCC_V25.pdf

[13] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on gammatone filters for robust speech recognition," in *Proc. Circuits and Systems (ISCAS)*, May 2013, pp. 305–308.

[14] M. Hansson-Sandsten and J. Sandberg, "Optimal cepstrum estimation using multiple windows," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 3077–3080.

[15] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. OShaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech Communication*, vol. 55, no. 2, pp. 237 – 251, 2013.

[16] D. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, Sept 1982.

[17] K. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," *Signal Processing, IEEE Transactions on*, vol. 43, no. 1, pp. 188–195, Jan 1995.

[18] M. Hansson and G. Salomonsson, "A multiple window method for estimation of peaked spectra," *Signal Processing, IEEE Transactions on*, vol. 45, no. 3, pp. 778–781, Mar 1997.

[19] R. D. Patterson, M. Unoki, and T. Irino, "Extending the domain of center frequencies for the compressive gammachirp auditory filter," *The Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1529–1542, 2003.

[20] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.